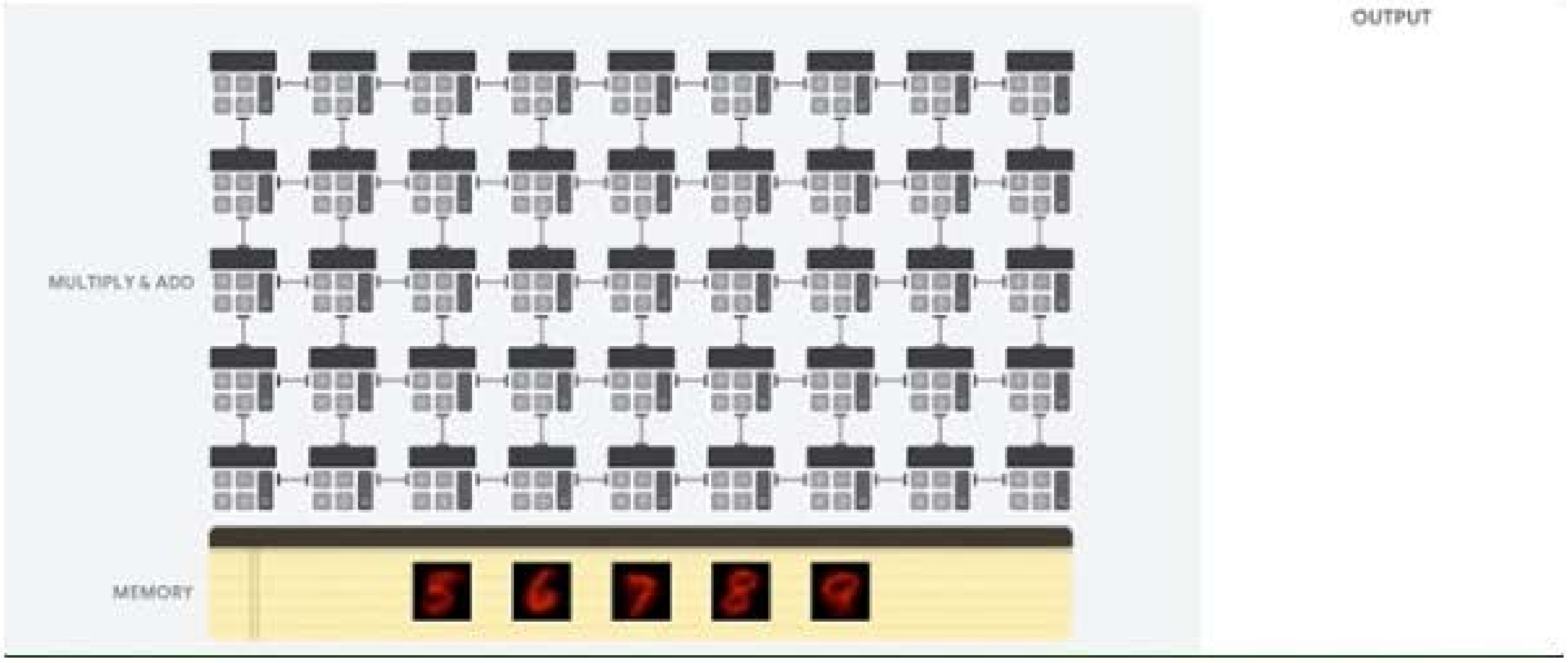
Exhibit C

How a TPU works

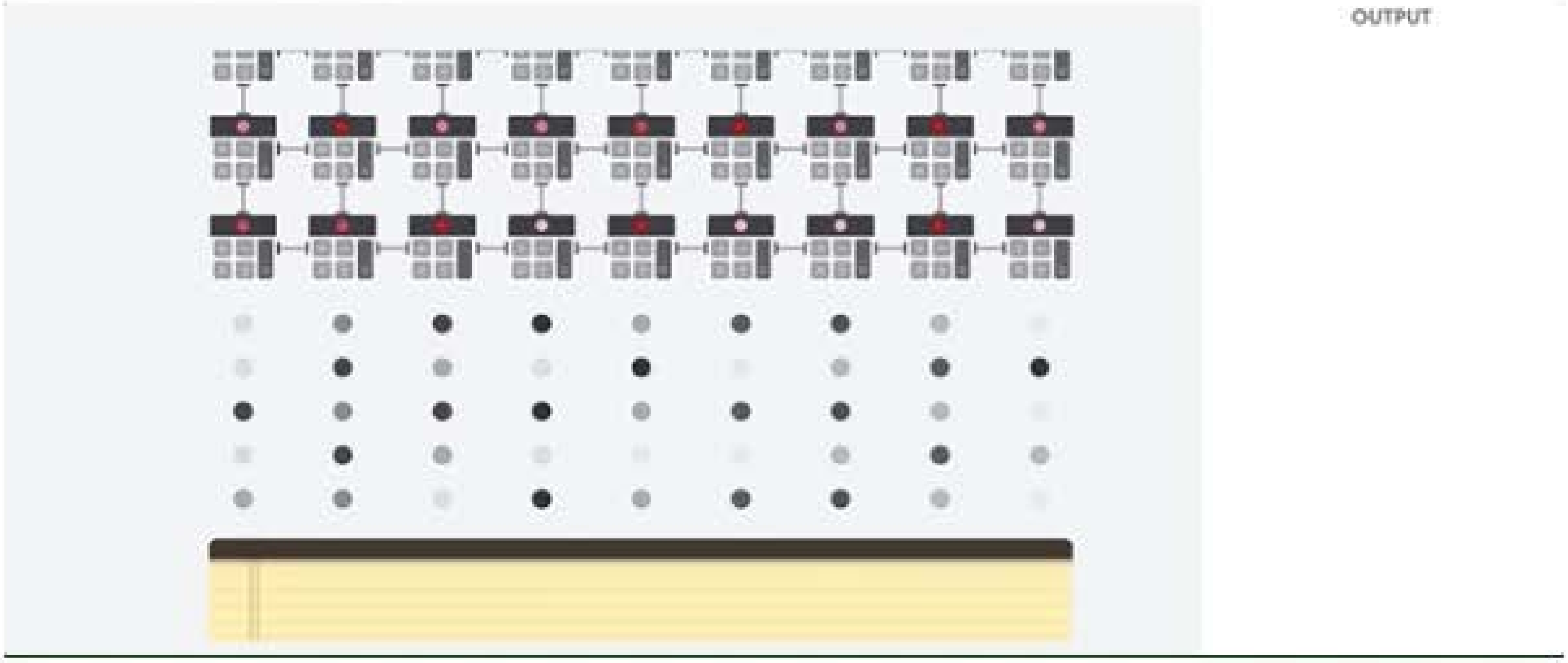
Google designed Cloud TPUs as a matrix processor specialized for neural network workloads. TPUs can't run word processors, control rocket engines, or execute bank transactions, but they can handle the massive multiplications and additions for neural networks at very fast speeds while consuming much less power and inside a smaller physical footprint.

One benefit TPUs have over other devices is a major reduction of the von Neumann bottleneck. Because the primary task for this processor is matrix processing, hardware designers of the TPU knew every calculation step to perform that operation. So they were able to place thousands of multipliers and adders and connect them to each other directly to form a large physical matrix of those operators. This is called a systolic array architecture. In the case of Cloud TPU v2, there are two systolic arrays of 128 x 128, aggregating 32,768 ALUs for 16 bit floating point values in a single processor.

Let's see how a systolic array executes the neural network calculations. At first, the TPU loads the parameters from memory into the matrix of multipliers and adders.



Then, the TPU loads data from memory. As each multiplication is executed, the result will be passed to the next multipliers while taking the summation at the same time. So the output will be the summation of all multiplication results between data and parameters. During the whole process of massive calculations and data passing, no memory access is required at all.



As a result, TPUs can achieve a high computational throughput on neural network calculations with much less power consumption and smaller footprint.